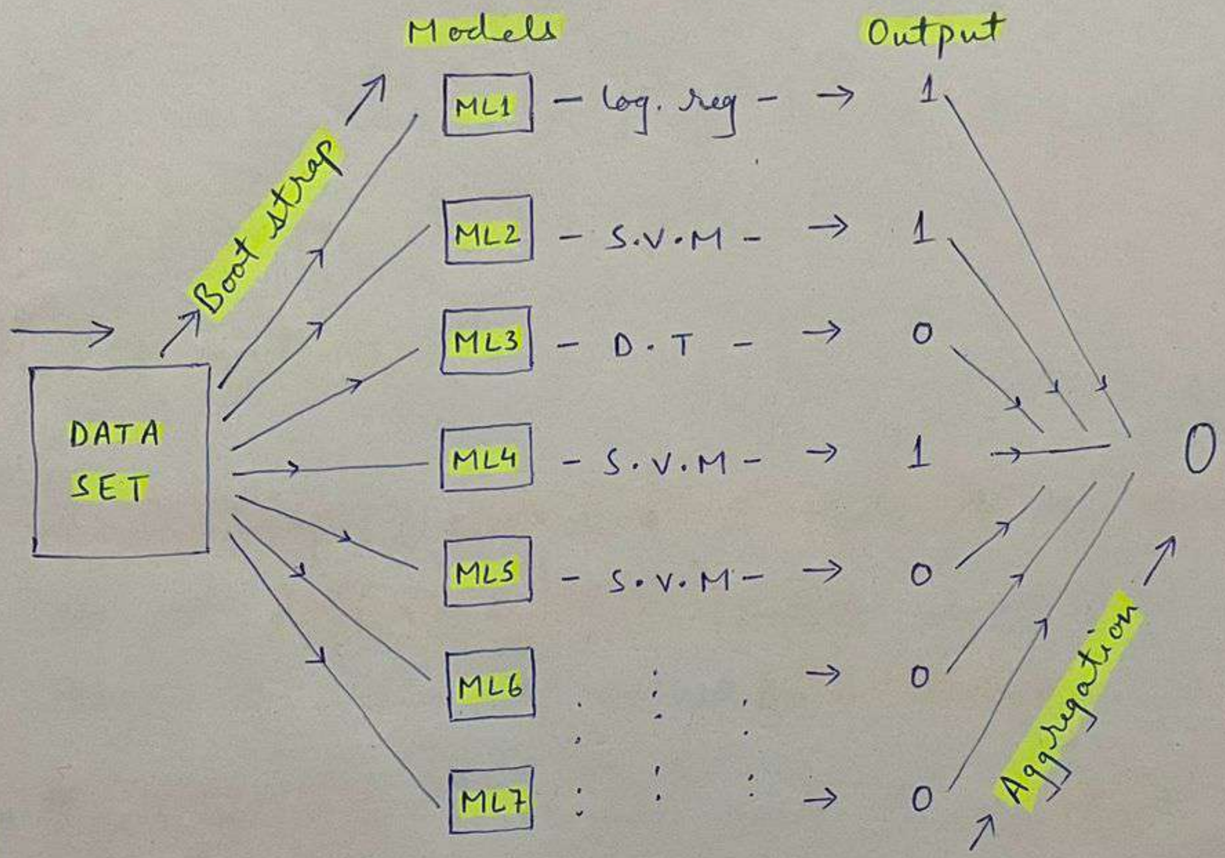
Random Forest

Bagging :



Consider we have dataset of classification problem and we perform set of operations using different algo(Algorithm) parallely to get corresponding output.

Once we have the output after operations from n different models we decide final output based on the majority voted classifier means what max of Algo has predicted. For above dataset output is 0 predicted by 4 Algo & 1 predicted by 3. Thus final output is 0.

Above concept of combining output of multiple Parallel individual Models to predict final output is called Bagging of Algarithm.

Two words comes in play during Bagging :

- Boot strap.
- Aggregation.

- **Boot strap** : Dividing of dataset to pass on to different ML. Models using different Algo to train is called Boot strap.

- **Aggregation** : Once we have the output from different trained model generated after Boot strap. Process of merging those output to get most occured final output is called Aggregation.

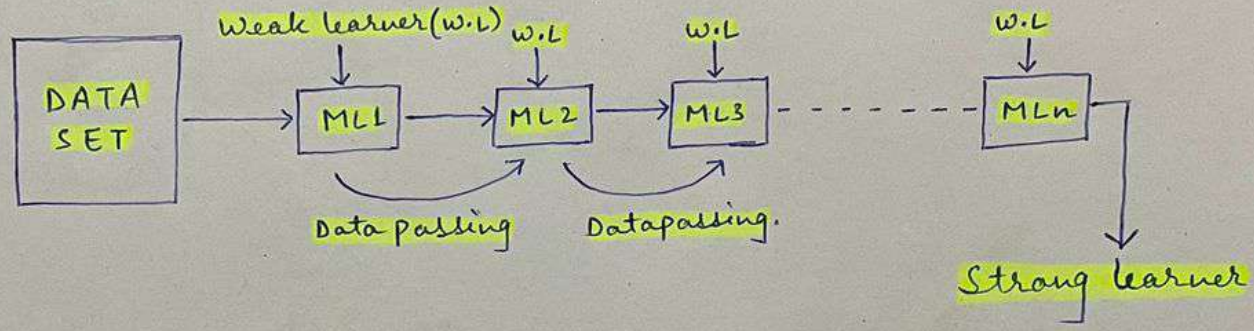
Technique of Bootstrap and Aggregation is called Ensemble technique.

Random forest classifier :
when we choose final Aggregated output based on Majar vating classifier means choosing the one having max. no. of output as final output.

Random forest regression:

when we choose final Aggregated output based on Average or Mean of the output of Bootstraped individual Algorithm.

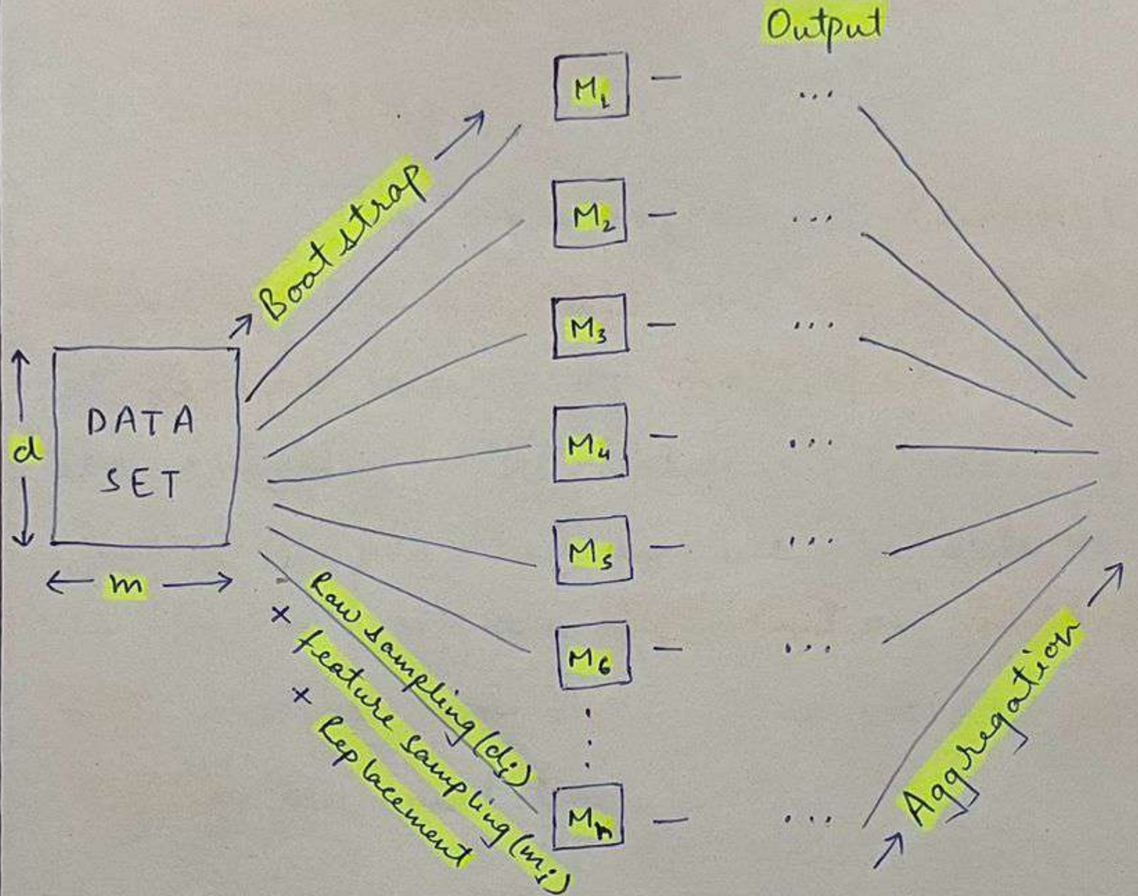
Boosting:



When we apply M.L algo on the dataset which passes through weak learner it continues to pass the data to going forward weak learner for which the Algo is not Performing well. For datas for which W.L are performing well will give the output.

All the M.L Algo's are connected sequentially which was earlier parallely in case of Bagging.

Random forest classification / Regression:



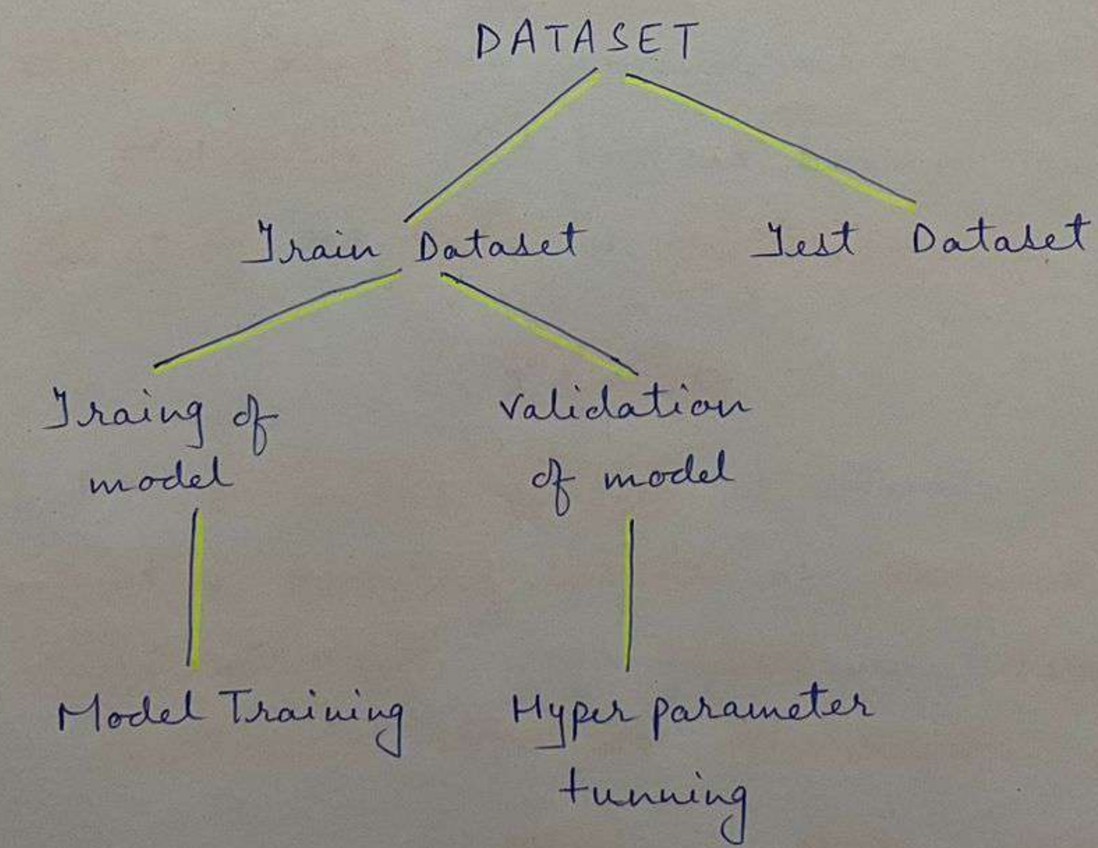
Consider parent data set d which will be passed to further ML models with different Algo $M_1, M_2, M_3 \ldots\ldots M_n$ with dataset splited as $d_1, d_2, d_3 \ldots. d_n$ and feature sampling splited as $m_1, m_2, m_3 \ldots\ldots m_n$.

There is possibility of same datapoints being available in different M.L Algo means same data in ($d_1 \rightarrow d_2 \rightarrow d_n$) few data points may collide.

when we draw Individual Decision Tree on huge dataset, there is maximum chance of overfitting.

Thus by using above technique of splitting the datasets and applying Algo over it reduces chance of overfitting. Thus we are reducing High variance in case of over-fitting to low variance. we are able to do so because we are dependent on multiple D.T rather than one. Thus we are generaliging the model.

Out of Bag Evaluation:



Oob - Score :

When we do prediction of the Validation data. Let's the accuracy comes 85%.

Then, Validation error

$$= 1 - .85$$

$$= .15$$

0.15 is validation error and is also called Oob - Score error.

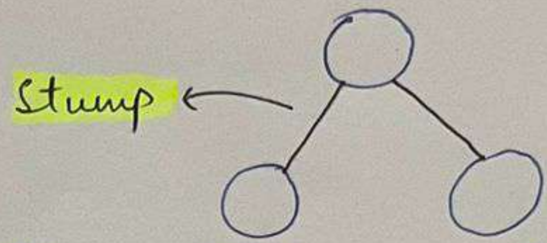
Validation error means what % of data out of whole data is Untouched means not being selected in consideration in any of M.L Algo. It happens because we are selecting the data on random basis.

Ada Boost.

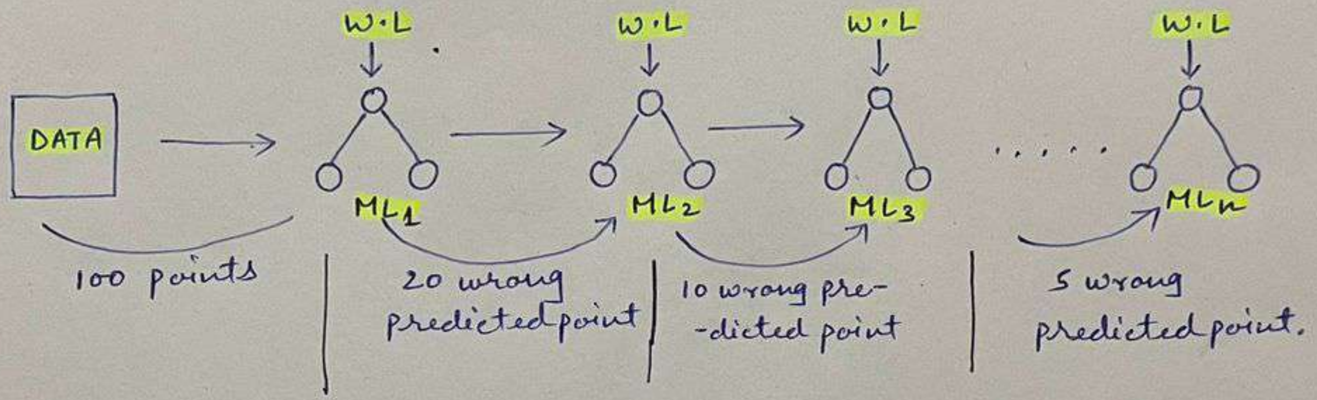
It's Ensemble method as stands for Adaptive Boosting. Weights are assigned to each instances, with higher weight assigned to incorrectly classified instances.

When we choose depth of D.T as 1, so less depth led to Underfitting.

Decision Tree (D.T) with just depth = 1 is called Stump



W.L : Weak learner.



Suppose initially we have 100 datapoints. Being passed to Model (ML_1) it predict right result for 80 points & wrong for 20. Then only wrong one will be passed to further Models [weak learns] and same will continue till we get right prediction for all points.

Final function:

$$= f = \alpha_1(ML_1) + \alpha_2(ML_2) + \alpha_3(ML_3) + \dots \dots \dots \dots + \alpha_n(ML_n)$$

α : weight

ML_i : Weak learner